Capstone Project: End-to-End Data Science Project

Project Overview:

As part of a growing e-commerce company, you have been tasked with analyzing product data to better understand market trends and customer preferences. Your objective is to leverage data science techniques to enhance the company's product offerings and marketing strategies. This project will involve collecting data, performing analysis, and using machine learning to derive actionable insights.

1. Web Scraping

You will begin by identifying an e-commerce website that allows for web scraping. Your goal is to extract essential product data, including:

- **Product Name**
- **Price**
- **Category**
- **Ratings**
- **Number of Reviews**

Using Python libraries such as Beautiful Soup or Scrapy, you will develop a script to scrape this data and save it in a CSV file. Ensure compliance with the website's terms of service to maintain legal and ethical standards.

2. Data Cleaning

Once you have your CSV file, the next step is data cleaning. You will load the data into a pandas DataFrame and perform the following tasks:

- Identify and handle missing values.
- Remove duplicates and irrelevant entries.
- Standardize formats (e.g., currency and text casing).
- Conduct exploratory data analysis (EDA) to uncover initial insights and data distributions.

3. Data Storage

After cleaning, you will set up a relational database to store your refined data. Using SQLAlchemy or similar libraries, you will push the cleaned dataset into the database for future access and analysis.

4. Unsupervised Learning

With your data securely stored, you will retrieve it for unsupervised learning. Your goal is to identify patterns and group similar products together. You will:

- Choose a suitable clustering algorithm (e.g., K-means).
- Experiment with different numbers of clusters (n) to find the optimal grouping.
- Add a new column to your dataset to indicate each product's cluster membership.

5. Supervised Learning

Next, you will shift to supervised learning. You will apply various classification algorithms to predict product categories or other relevant labels based on features. The algorithms to implement include:

- Logistic Regression
- Support Vector Machine (SVM)
- k-Nearest Neighbors (k-NN)
- Random Forest
- XGBoost

You will evaluate each model's performance using metrics such as accuracy and F1 score to determine which algorithm provides the best results.

6. Hyperparameter Tuning

After identifying the best-performing model, you will optimize its performance through hyperparameter tuning. Utilizing techniques like Grid Search or Random Search, you will experiment with different parameters and document the best configuration that enhances the model's accuracy.

7. Documentation and Submission

Finally, you will compile your findings and processes into a comprehensive project report. This documentation will include:

- A detailed methodology outlining each step of your analysis.
- Insights gained from the data and modeling processes.
- A PowerPoint presentation summarizing your project for stakeholders.
- Well-organized and commented code files showcasing your work.

Evaluation Rubric

Evaluation Criteria	Marks
1. Web Scraping	10
2. Data Cleaning	15
3. Data Storage	10
4. Unsupervised Learning (Clustering)	10
5. Supervised Learning (Classification)	15
6. Hyperparameter Tuning	10
7. Documentation and Reporting	10
8. Presentation	20
Total	**100**

Expected Outcomes

Through this project, you will develop a deeper understanding of data collection, cleaning, analysis, and machine learning techniques. Your insights will not only help the e-commerce company refine its product offerings but also demonstrate your ability to apply data science principles in a real-world scenario.

Good luck with your project! And remember to document each step carefully to highlight your learning journey!